A New Approach to the Fundamental Problem of Applied Statistics

by

Seymour Geisser*

University of Minnesota

Technical Report No. 235

*Research partially supported by US Army grant DAHC04-74-G-0216.

A NEW APPROACH TO THE FUNDAMENTAL PROBLEM OF APPLIED STATISTICS

by Seymour Geisser

University of Minnesota

1. Introduction

The prediction of future binary trials when a set of such similar trials have alrady been observed was formally denoted as the fundamental problem of Applied Statistics by Karl Pearson shortly after the turn of the century. Undoubtedly its importance was perceived, if not articulated, long before as it is the basic paradigm of inductive inference, parametric models notwithstanding. In a very real sense, statistical theory should be directed towards providing methods regarding situations such as: Having observed N binary events of $N+M$ under scrutiny it is required to make some sensible determination of the number of the remaining M that are of each kind success or failure, say. For such a situation, vague as it is, there are many possible statistical models that may be applicable, the most common being a sequence of independent Bernoulli trials each having the same chance of success.

We consider the most primitive and basic case which is essentially an urn containing $N+M$ balls some marked 1 for success and others marked 0 for failure, the numbers of each are unknown. N of them are drawn from the urn in some manner, perhaps at random, perhaps haphazardly, or even purposively, we do not necessarily know but we are compelled to guess at the composition of the remaining balls. Unless there is reason to believe they were drawn in some biased manner that we could take advantage of, it would seem reasonable that we might as well assume that the draw

was such that every set of N balls out of $N+M$ has the same chance of appearing. As to how the $N+M$ balls got into the urn or became the subject of our purview, we do not necessarily know although we may have some suspicion. They may have been generated by some physical mechanism say the toss of the same fair or biased coin or by a series of tosses with coins with differing biases or they may just have been there, as it were.

The approach used here is termed the predictive sample reuse method, Geisser (1974) or the cross-validatory method, Stone (1974). Briefly, as it is applied here, it possesses four ingredients: A set of N observations $X = (x_1,\ldots,x_N)$; a predictive function $f(X;\alpha)$; a schema $\mathcal{S}$ of observational omissions; an average discrepancy measure $D_{N,1}(\alpha)$. Actually, we shall be using throughout this paper a schema of single observational omissions, so that we need not discuss this aspect further. A predictive function $f(X;\alpha)$ is selected depending on some unknown value α and then is given its final form denoted by $\hat{f}(X;\alpha) = f(X,\hat{\alpha}) = \hat{f}$ by insertion of a suitable value for α, say $\hat{\alpha}$. In order to find an appropriate value for α an average discrepancy measure is chosen

$$D_{N,1}(\alpha) = \frac{1}{N}\sum_{j=1}^{N} d(f(X^{(j)},\alpha),g(x_j)) \tag{1.1}$$

where d is the discrepancy between $g(x_j)$, a function of the omitted observation, and $f(X^{(j)},\alpha)$ the predictor of $g(x_j)$; with $X^{(j)}$ defined as the original set X with x_j deleted; the subscripts of $D_{N,1}(\alpha)$ refer to the number of values observed, say N, and the number omitted which throughout this paper is 1. The next step is the minimization of $D_{N,1}(\alpha)$ with respect to $\alpha \in \Omega$. When this yields a unique value $\hat{\alpha}$, the predictor then is $f(X,\hat{\alpha}) = \hat{f}$.

Basically then, this is a distribution-free, empirical method that simulates the predictive process as best it can given the lack of specificity

involved. In a sense the only modeling in this paradigm is via the form for the predictive function. Even this is rather modest as no intrinsic physical meaning is necessarily attached to a predictor per se, but a variety of predictive functions may be tried out and compared via cross-validatory assessments. The one, most apparently suitable for the given set of data, then is given prime consideration.

For the problem at hand there are three predictive forms that have drawn the most attention for predicting the number of future successes. The classical one is Mr/N, r being the number of successes, when the underlying model entertained is that $x_1,\ldots,x_N$ are Bernoulli iid variates with unknown probability of success θ. If in addition a Beta prior distribution for θ with known parameters g, the prior expectation of success, and α, a reflection of the prior precision of g, is utilized, then the predictor is $M(r+g\alpha)(N+\alpha)^{-1}$. A third is the predictor $M(\alpha rN^{-1}+(1-\alpha)g)$ derived from the decision-theoretic frequentist approach of Stein (1956). In most previous work the emphasis has been on the estimation of the parametric probability of success θ which could be covered by setting $M=1$ here. A variety of work based either on Bayesian or frequentist approaches or combinations thereof utilizing estimators of this kind and extended to multinomial situations can be cited; Fienberg and Holland (1974), Good (1965), Sutherland et al (1974).

In what follows we shall consider these three predictive functions, remembering of course that the first needs no further elucidation being independent of α. For the other two we require that g be guessed a priori and that a suitable α be found from the data itself. Further, we shall concern ourselves with only two measures of discrepancy, the squared and absolute difference between predictor and predictand using the notation $s^2_{N,1}(\alpha)$ and $S_{N,1}(\alpha)$ for $D_{N,1}(\alpha)$ respectively.

2. Squared Discrepancy

We now investigate the two forms alternative to the classical predictor discussed in the previous section as predictive functions in the paradigm under study. For $0 \geq g \geq 1$, $N > 1$,

$$f_1 = M(\frac{r+g\alpha}{N+\alpha}) \qquad \alpha \geq 0 , \tag{2.1}$$

$$f_2 = M(\alpha\frac{r}{N} + (1-\alpha)g) \qquad 0 \leq \alpha \leq 1 \tag{2.2}$$

Both of these were given some attention recently from the predictive sample reuse viewpoint, Geisser (1975), especially f_1.

Firstly, we shall apply a schema of one-at-a-time omissions and squared discrepancy using both predictive functions. For f_1 , the average squared discrepancy is

$$s^2_{N,1}(\alpha) = M^2N^{-1} \sum_{j=1}^{N} (\frac{r-x_j+g\alpha}{N-1+\alpha} - x_j)^2$$

$$= M^2N^{-1}(N-1+\alpha)^{-2}[(N-r)(r+g\alpha)^2 + r(N-r+\alpha(1-g))^2] . \tag{2.3}$$

Minimization of the above with respect to α yields

$$\left.\begin{aligned} \hat{\alpha} &= \frac{r(N-r)}{(N-1)[g^2(N-r) + (1-g)^2r] - r(N-r)} \quad \text{if } (N-1)[g^2(N-r) + (1-g)^2r] > r(N-r) \\ \hat{\alpha} &= \infty \qquad\qquad \text{otherwise.} \end{aligned}\right\} \tag{2.4}$$

Hence for all appropriate g , $\hat{f}_{12} = Mg$ (the subscript 2 indicates squared discrepancy), whenever

$$\frac{r}{N} \in g(1-N^{-1}) + (2N)^{-1} \pm (2N)^{-1}(4g(1-g)(N-1)+1)^{\frac{1}{2}} \tag{2.5}$$

We note that the interval is not quite symmetric about g but its

center is pulled slightly towards $\frac{1}{2}$ but this is sensible due to the finite range and lack of symmetry for the distribution of rN^{-1} (if a probability model on the observations were imposed) and $g \neq \frac{1}{2}$ were the true probability of a success. As N increases the interval tends to $g \pm (g(1-g)/N)^{\frac{1}{2}}$ which is approximately plus or minus one standard deviation from g if $x_1,\ldots,x_N$ were considered to be iid Bernoulli variates with g being the true chance of a success. Also for r/N outside the interval $\hat{f}_{12} \to MrN^{-1}$ as N increases. We also observe that

$$|\hat{f}_{12} - Mg| = \frac{N}{N + \hat{\alpha}} \; |MrN^{-1} - Mg| \tag{2.6}$$

so that we define $N^{-1}(N + \hat{\alpha}) = C_{12}(g) \geq 1$ as the compression ratio. This is basically a measure of how much closer $\hat{f}_{12}$ is to Mg as opposed to the usual predictor MrN^{-1}. The higher the compression ratio the more $\hat{f}_{12}$ is pulled towards Mg.

For f_2, the average squared discrepancy is

$$M^2 N^{-1} \sum_{j=1}^{N} \left[\frac{\alpha(r-x_j)}{N-1} + (1 - \alpha)g - x_j\right]^2 . \tag{2.7}$$

We note a general solution for combining a mean (r/N in this case) and a prior estimate g has already been provided by Geisser (1975). Applying the results there to this special case yields

$$\left.\begin{aligned} \hat{\alpha} &= \frac{N(g-rN^{-1})^2 - (N-1)^{-1}r(1-rN^{-1})}{N(g-rN^{-1})^2 + (n-1)^{-2}r(1-rN^{-1})} \quad \text{if } r(1-rN^{-1}) < (N-1)N(g-rN^{-1})^2 \\ \hat{\alpha} &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise .} \end{aligned}\right\} \tag{2.8}$$

Hence $\hat{f}_{22} = Mg$ whenever

$$\frac{r}{N} \in g(1-N^{-1}) + (2N)^{-1} \pm (2N)^{-1}(4g(1-g)(N-1) + 1)^{\frac{1}{2}} \tag{2.9}$$

which is identical to $\hat{f}_{12}$ in this range. Outside this range they differ

slightly, though they share the same asymptotic properties. Here

$$|\hat{f}_{22}-Mg| = \hat{\alpha}\ |MrN^{-1}-Mg|$$

so that the compression ratio is $\hat{\alpha}^{-1} = C_{22}(g) \geq 1$. It is clear that for every fixed g , both $C_{22}(g)$ and $C_{12}(g)$ tend to unity as N increases. Further, it can easily be shown algebraically that

$$C_{12}(g) = \frac{(N-1)(r-gN)^2}{(N-1)(r-gN)^2 - r(N-r)} \tag{2.10}$$

and

$$C_{22}(g) = C_{12}(g) + \frac{(N-1)^{-2}r(1-rN^{-1})}{N(g-rN^{-1})^2 - (N-1)^{-1}r(1-rN^{-1})} \quad . \tag{2.11}$$

Thus $C_{22}(g) > C_{12}(g)$ whenever the denominator of the second term of the right hand side is greater than zero and $r \neq 0$ or N , otherwise $C_{22}(g) \equiv C_{12}(g)$. This of course means that $\hat{f}_{22}$ always is endowed with at least as great a compression ratio as $\hat{f}_{12}$. Since this holds for all g this includes the central case $g = \frac{1}{2}$ as well. The central case for $\hat{f}_{12}$ was also given for multiple omissions in Geisser (1975). For the one-at-a-time omission schema we obtain

$$\left.\begin{aligned} \hat{f}_{12} &= \frac{Mr(N+1-2r)}{(N-1)(N-2r)} \qquad \text{if} \quad N(N-1) > 4r(N-r) \\ &= \frac{M}{2} \qquad\qquad\qquad \text{otherwise.} \end{aligned}\right\} \tag{2.11}$$

3. Absolute Discrepancy

We now consider the application of absolute discrepancy to this prediction problem. Thus for f_1 we would be required to minimize

$$S_{N,1}(\alpha) = N^{-1}M \sum_{j=1}^{N} \left| \frac{r-x_j+g\alpha}{N-1+\alpha} - x_j \right| = N^{-1}M(N-1+\alpha)^{-1}[2r(N-r)+\alpha[g(N-2r)+r]] \quad (3.1)$$

with respect to $\alpha \geq 0$. Minimization yields

$$\left.\begin{aligned} &\hat{\alpha} = 0 \ , \ \hat{f}_{11} = MrN^{-1} \quad \text{if } (N-1)[g(N-2r)+r] \geq 2r(N-r) \\ &\hat{\alpha} = \infty \ , \ \hat{f}_{11} = Mg \quad \text{otherwise} \end{aligned}\right\} \quad (3.2)$$

where the second subscript indicates absolute discrepancy. This result has been alluded to previously by Geisser (1975), and discussed in some detail for $g = \frac{1}{2}$, the central case. For this special case we note that

$$\left.\begin{aligned} &\hat{f}_{11} = M/2 \quad \text{if} \quad r/N \in \tfrac{1}{2} \pm \tfrac{1}{2}\sqrt{N} \\ &\hat{f}_{11} = Mr/N \quad \text{otherwise} \end{aligned}\right\} \quad (3.3)$$

which is eminently sensible. However for $g \neq \frac{1}{2}$ some difficulties arise since the procedure requires that we predict Mg whenever r is in the interval that includes the roots of the quadratic equation

$$2r^2 - [N+1+2g(N-1)]r + gN(N-1) = 0 \ . \quad (3.4)$$

That this result is not very sensible arises from the fact that the roots of the above quadratic equation tend to $N/2$ and gN as N increases. Hence no matter how large N we must perforce predict Mg if r/N falls between g and $\frac{1}{2}$, no matter what g is assumed. This lack of "consistency"

is not necessarily the fault of either the predictive function or the method but is largely due to the discrepancy measure as we will shortly observe.

Now consider applying absolute discrepancy to f_2, so that we need minimize

$$S_{N,1}(\alpha) = MN^{-1}\sum_{j=1}^{N}\left|\frac{\alpha(r-x_j)}{N-1} + (1-\alpha)g - x_j\right| \tag{3.5}$$

$$= MN^{-1}(N-1)^{-1}\{(N-1)[g(N-2r)+r] + \alpha[(N-2r)(r-g(N-1))+r]\} .$$

Hence minimization of the above discrepancy with respct to α leads to

$$\left.\begin{aligned} &\hat{\alpha} = 1 , \quad \hat{f}_{21} = Mr/N \quad \text{if} \quad r \le (N-2r)(g(N-1)-r) \\ &\hat{\alpha} = 0 , \quad \hat{f}_{21} = Mg \qquad \text{otherwise} . \end{aligned}\right\} \tag{3.6}$$

A little algebra shows that the condition $r \ge (g(N-1)-r)(N-2r)$ is equivalent to the quadratic equation (3.4) given in the previous case. Hence the method yields the same predictor for both predictive functions when absolute discrepancy is utilized. Actually the algebra here is transparent as to what is occurring in regard to the discrepancy function. Note for example, if $r = N/2$ the absolute discrepancy has a minimum at $\alpha = 0$ irrespective of g or N.

4. A cross-validatory comparison

We note that for all g, both predictive functions yield the same predictor when using absolute discrepancy although this does not hold for squared discrepancy. This common result we denote for $g = \frac{1}{2}$ as

$$\begin{aligned} \hat{f} &= Mr/N && \text{for } (N-1)N \geq 4r(N-r) \\ \hat{f} &= M/2 && \text{otherwise} \end{aligned} \tag{4.1}$$

Suppose we wished to assess this predictor as against the natural predictor Mr/N for any particular r. One way of comparing two predictors within this frame of reference is to make a cross-validatory assessment of them for the data set in hand, see Stone (1974), or Geisser (1974, 1975). This is accomplished by putting aside an observation x_j, then on the remaining $N-1$ observations to calculate the requisite predictor $f(X^{(j)},\hat{\alpha}_j)$ as if x_j did not exist and then repeating this for each j and further calculating an average discrepancy

$$N^{-1} \sum_{j=1}^{N} d(f(X^{(j)},\hat{\alpha}_j),Mx_j) = D^*_{N-1}$$

where $d(fX^{(j)}, \hat{\alpha}_j), Mx_j)$ is some

defined discrepancy between $f(X^{(j)},\hat{\alpha}_j)$ and Mx_j. Suppose we were to let $d(f(X^{(j)},\hat{\alpha}_j),Mx_j)=|f(X^{(j)},\hat{\alpha}_j) - Mx_j|$, the absolute discrepancy. Then it is possible in this case, as it is usually not, to calculate D^*_{N-1} explicitedly, which is here now denoted as S^*_{N-1}, for any value of r for the above procedure. Thus after some calculation we obtain

$$\left.\begin{aligned} S^*_{N-1} &= \frac{2r(N-r)M}{N(N-1)} && \text{if } \mathrm{Max}[4r(N-1-r),4(r-1)(N-r)] < (N-1)(N-2) \\ &= M/2 && \text{if } \mathrm{Min}[4r(N-1-r),4(r-1)(N-r)] > (N-1)(N-2) \\ &= \frac{r(3N-2r-1)M}{2N(N-1)} && \text{if } 4r(N-1-r) \leq (N-1)(N-2) \leq 4(r-1)(N-r) \\ &= \frac{(N-r)(N+2r-1)M}{2N(N-1)} && \text{if } 4(r-1)(N-r) \leq (N-1)(N-2) \leq 4(r-1)(N-1-r) \end{aligned}\right\} \tag{4.2}$$

Further, for the classical predictor Mr/N it is easy to show that $S^*_{N-1} = 2r(N-r)M/N(N-1)$ for all r. The following inequalities also obtain;

$$\frac{2r(N-r)}{N(N-1)} > \tfrac{1}{2} \quad \text{if } \mathrm{Min}[4r(N-1-r),\ 4(r-1)(N-r)] > (N-1)(N-2) \tag{4.3}$$

but

$$\tfrac{1}{2} < \frac{2r(N-r)}{N(N-1)} < \frac{r[3N-2r-1]}{2N(N-1)} \quad \text{if} \quad 4r(N-1-r) \le (N-1)(N-2) \le 4(r-1)(N-r) \tag{4.4}$$

and

$$\tfrac{1}{2} < \frac{2r(N-r)}{N(N-1)} < \frac{(N-r)(N+2r-1)}{2N(N-1)} \quad \text{if } 4(r-1)(N-r) \le (N-1)(N-2) \le 4r(N-1-r) \tag{4.5}$$

Now by virtue of (4.2), for any particular N and some r_o depending on N it can be shown that for all r such that $r < r_o$ and $r > N-r_o$ both assessments are the same. Further, for at most the four points $r = r_o$, r_o+1, $N-r_o$, $N-r_o-1$, and at least for two of these the usual predictor is better by virtue of (4.4) and (4.5). By the same token and (4.3), $\hat{f}$ is superior for all r such that $r_o+1 < r < N-r_o-1$. Hence, except for at most four values of r and possibly only two, $\hat{f}$ always dominates the usual predictor on a cross-validatory assessment depending on absolute dixcrepancy.

We can also compare the two previous methods with $\hat{f}_{12}$ when $g = \frac{1}{2}$. In this case as we have noted before in (2.11)

$$\hat{f}_{12} = \frac{Mr(N+1-2r)}{(N-1)(N-2r)} \qquad \text{if} \quad N(N-1) > 4r(N-r)$$

$$\hat{f}_{12} = M/2 \qquad \text{otherwise}.$$

Clearly, $\hat{f}_{12}$ is pulled at least as much if not more to $M/2$ as $\hat{f}$. Again we can compute S^*_{N-1} for this procedure and here we obtain

$$
\left.
\begin{aligned}
S^*_{N-1} &= \frac{2r(N-r)M(N-2r)^2}{N(N-2)(N-1-2r)(N+1-2r)} \quad \text{if } \mathrm{Max}[4r(N-1-r), 4(r-1)(N-r)] < (N-1)(N-2) \\
&= M/2 \quad \text{if } \mathrm{Min}[4r(N-1-r),\ 4(r-1)(N-r)] > (N-1)(N-2) \\
&= \frac{rM[2(N-r)(N-2r) + (N-2)(N-1-2r)]}{2N(N-2)(N-1-2r)} \quad \text{if } 4r(N-1-r) \le (N-1)(N-2) \le 4(r-1)(N-r) \\
&= \frac{(N-r)M[2r(N-2r) + (N-2)(N+1-2r)]}{2N(N-2)(N+1-2r)} \quad \text{if } 4(r-1)(N-r) \le (N-1)(N-2) \le 4r(N-1-r)
\end{aligned}
\right\} \quad (4.6)
$$

By comparing (4.6) with (4.2) we observe that the cross-validatory assessment of $\hat{f}$ dominates that of $\hat{f}_{12}$ for all r while the usual predictor is sometimes better and sometimes worse than $\hat{f}_{12}$ depending on r. Therefore, a cross-validatory assessment using absolute discrepancy clearly implies preference for $\hat{f}$ as opposed to $\hat{f}_{12}$. Of course a cross-validatory assessment using squared discrepancy would undoubtedly completely change the dominance structure. In view of this it may be wise to use several different cross-validatory assessments if no one of them is inherently compelling as the criterion before selecting a particular predictor. What is certainly desirable is explicit formulae for a variety of appropriate discrepancy measures so that the choice of the predictor can be made conveniently before hand without a great deal of heavy computation. The most informative comparison is a graphical display of the raw difference $f(X^{(j)},\hat{\alpha}_j) - Mx_j$ for the various predictors either as a histogram or as an empirical distribution function, but this requires the full computation.

5. Preferenctial Predictive Sets

In the previous discussion we asserted that what was at issue was the prediction of the number of future or as yet unobserved "successes" in a paradigm where only N out of $N+M$ observables had been determined. The previous solutions presented for $\hat{f}$ were not necessarily integers. Ostensibly we should predict at least a single integer or perhaps several integers. Now for a Bayesian this presents no difficulty whatever (excepting of course the high structure assumptions involved) since he can derive the predictive distribution of the sum of the M unobserved values, say, $x_{N+1}+\ldots+X_{N+M}= z$, where $z = 0,1,\ldots,M$ with associated predictive probability $p(z)$. Hence he would be in a position to obtain a single set of integers I and a predictive probability that $z \in I$ which of course is the ultimate for predictive purposes. If the situation were such that he is allowed to guess, say, C out of the $M+1$ possible integral values, he could choose a set of C which yielded the maximum predictive probability provided there was a unique set. In any event because we are dealing with a low structure paradigm we cannot presume to obtain such a fine inferential yield.

In our case, if a single integer is desired, we could predict the integer closest to $\hat{f}$ or better choose that integer which minimized $D_{N,1}(\alpha)$ for values of f restricted to $0,1,\ldots,M$. It is clear that neither approach need yield a unique integer since $\hat{f}$ may be the mid-point between two integers in the first case or analogously there may be more than one integral minimum in the second case. When this occurs and also for other reasons it seems natural to expand our predictor to more than one integer value. On the other hand, if we were initially given greater lattitude and could include $C = 1,\ldots,M$ integral choices out of the $M+1$ possible predictive values (obviously $C = M+1$ is trivial) then the question at issue is which C out of $M+1$ should we select. We shall list

three possible ways of resolving this problem, none of which will be necessarilywholly adequate in any particular situation.

One possible resolution of this problem is to calculate the average discrepancy,

$$D_{N,1}(\alpha) = N^{-1} \sum_{j=1}^{N} d(f(X^{(j)},\alpha),Mx_j), \quad \alpha \in \Omega$$

for those values α which are solutions of $f(X,\alpha) = j$, $j=0,\ldots,M$. for the given range on α. Then one might prefer $t = f(X,\alpha_t)$ to $k = f(X,\alpha_k)$ if $D_{N,1}(\alpha_t) < D_{N,1}(\alpha_k)$. If a set of C were required one could choose those C of the set of $\{f(X,\alpha)\}$ that were derived from the smallest set of corresponding $\{D_{N,1}(\alpha)\}$. That this is not entirely adequate can be immediately inferred from the fact that for $g = \frac{1}{2}$ and $N = 2r$, $f(X,\alpha) \equiv M/2$ for all α. Hence whenever $f(X,\alpha)$ has a very restricted domain, this procedure may not be very useful. Its major advantage lies in the fact that no additional assumptions are required.

Another possible method follows from the intuitive prescription of including $C-1$ integers adjacent to the initial "best" predictor. This is perhaps most conveniently accomplished, by adding and subtracting a multiple of the cross-validatory assessment discrepancy (using the same discrepancy measure that generated the predictor) that just included the requisite number of integers. For large sample sizes and some, perhaps not so modest additional assumptions this might also be translated into a predictive set with some approximate confidence coefficient attached to it. This procedure of course tacitly assumes that predictive values are graded preferentially in accordance with their proximity to the initial best value.

A third way of looking at the problem is to insert slightly more into the

initial structure by assuming that prior to observation one can order values of g. This could involve a preferential guessing of the number of successes out of the M unobserved results. One guesses what he, a priori, believes to be the best C out of M+1 guesses say $j_1,\ldots,j_C$ where $\{j_1,\ldots,j_C,\}$ is subset of the integers $\{0,1,\ldots,M+1\}$. Then the procedure is recomputed anew for each $g_i = j_i/M$ $i = 1,\ldots,C$ yielding a preferred set of C predictors $f^{(1)},\ldots,f^{(C)}$. These then may be each adjusted to nearest integers or best integers as in the first method. If the values $f^{(1)},\ldots,f^{(C)}$ are not distinct, one could continue the process i.e. using further preferred values of g until C distinct values result. By the same token one could obtain C distinct values using a subset of $\{j_1,\ldots,j_C\}$. This indicates the procedure should be undertaken stepwise.

This method, assuming more prior knowledge than originally specified, also allows some schizoid guessing. For example, it may be felt that a priori the success ratio should be either rather modest or fairly high with central values originally being unlikely for one reason or another.

Of course these three methods may obviously be extended to cases where we wish to predict not sets of integers but, say, intervals or regions. Here potential observations are essentially assumed to be continuous. For the first alternative we would fix the ratio

$$\frac{D_{N,1}(\alpha)}{D_{N,1}(\hat{\alpha})} \leq K \qquad \text{where } K \geq 1 \tag{5.1}$$

solving for all α satisfying the above and generating predictive regions of level K obtained from $f(X,\alpha)$ for all α satisfying (5.1).

The extension of the second method is of course obvious. The adaption of the third method to the continuous situation is somewhat more involved in its details and execution but not in its conception. However, we shall leave that for another discussion.

6. An Illustration

Suppose there is an urn containing 30 red and green balls about which we have an initial suspcion that the process by which they got into the urn was such that the number of each should not have been too far from equal. We draw 20 balls at random and find 8 are red and 12 green. What should we guess to be the number of red balls remaining in the urn? As an illustration of the previous work we shall use $\hat{f}_{12}$ as our predictor. Using $g = \frac{1}{2}$ as our prior guess we obtain $\hat{f}_{12} = 5$. Suppose we are allowed to predict three integers for the number of red balls remaining. Now

$$f_1 = 10\left(\frac{8 + \frac{1}{2}\alpha}{20 + \alpha}\right) \qquad \alpha \geq 0 \qquad (6.1)$$

implies only two integral solutions for f_1, namely $f_1=4$, if $\alpha = 0$ and $f_1 = 5$ if $\alpha = \infty$. Hence the first method can never yield more than these two integers and is deficient in this respect. The second method readily yields $f_1 = 4,5,6$. The third method requires we add some more initial guesses. Given our initial suspicion of roughly equal amounts so that our first guess was $g = \frac{5}{10}$, we would be inclined to produce second preferential guesses of $10g = 4$ and $10g = 6$. Calculating $f_{12}(\hat{\alpha},g)$ for the next two guess $g = .4$ and $g = .6$ yields a predictive set of three integers (4,5,6) which turns out to be no different than our initial guesses.

7. Remarks

The process generating the predictors presented here are "best" in terms of their definitions and frame of rererence. For a strict Bayesian who takes upon himself the whole yoke of assumptions that harness his structure, these predictors will in general turn out to be incoherent and would have only marginal interest e.g. as a good initial approximation for a Bayesian predictor that is analytically intractable. An example of this is given by Geisser (1974). However, in many real problems the most confirmed subjective Bayesian will often have serious doubts as to the feasibility of judiciously and fully executing the whole tightly structured process. He then may find these procedures usefully robust. He may reasonably regard inference as a serious affair and though coherency desirable, betting merely a game.

As for the classical statistician willing only at most to specify likelihoods of observations and primarily interested in the frequency property of estimators, they could regard these as such and are entitled to judge them accordingly. This would include a check on their admissibility, risk function properties and robustness for varying assumptions on the distribution of the observations. It is anticipated that the methods that flow from the predictive sample reuse approach including the particular ones developed here, will enjoy many of the frequency properties associated with good estimators.

Be all that as it may, it is worthwhile emphasizing that the approach taken here differs in attitude and intent regarding the analysis of data than the stringent classical and Bayesian views. Its intent is always prediction and it takes a much more relaxed and flexible attitude toward models. Although both of the former approaches can be directed towards

prediction, their central concern has always been parametric models and their estimation. If the requirements of those approaches are taken too seriously the statistician is so constrained that it is possible for his inference to bear little relation to the data at hand or to reality. We believe that statisticians involved in data analysis do take a much more flexible view towards the process of inference than are permitted by those approaches. This should be recognized by theoreticians and the predictive sample reuse view is a step in this direction.

References

Fienberg, S. and Holland, P. (1973), Simultaneous estimation of multinomial cell probabilites. JASA, Vol. 68, 683-691.

Geisser, S. (1974). A predictive approach to the random effect model. Biometrika, 61, 1, pp. 101-107.

Geisser, S. (1975). The predictive sample reuse method with applications. To appear in JASA.

Good, I.J. (1965). The Estimation of Probabilities, Cambridge, Mass., Massachusetts Institute of Technology Press.

Stein, C.M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. 3rd Berkeley Symposium, 1, pp. 197-206.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. JRSS, Series B, Vol. 36 (to appear).

Sutherland, M., Holland, P.W., and Fienberg, S.E., (1974). Combining Bayes and frequency approaches to estimate a multinomial parameter, in S.E. Fienberg and A. Zellner, eds., Studies in Bayesian Econometrics and Statistics, Amsterdam: North Holland Publishing Co.